Grapheme Based Speech Synthesis and Speech Recognition

S.P. Kishore
skishore@cs.cmu.edu

Carnegie Mellon University
&
International Institute of Information Technology
Focus Papers

• Alan W Black and Ariadana Font Llitjos, “Unit Selection With out Phoneme Set”, IEEE TTS Workshop 2002, Santa Monica, CA

Speech Synthesis without Phoneme Set

• Why without phoneme set?
• Increasing need for speech synthesis and speech recognition in new unsupported languages.
  – Availability of less phonetic knowledge in the new languages
  – Researchers/developers may not have (or access to) language expertise
  – Native speakers may not be consciously aware of the phonetic knowledge

• How hard/easy for well studied languages? –
  – An appropriate phoneme set for a well studied language may not be easy.
  – Orthography and phonetics may not have one to one relationship
Experiment in Spanish language

• Nature of the language: Written system and phonology is relatively close, but not one-to-one
• Letter set as phoneme set
  – 26 standard English letters
  – Accented characters a’, e’, i’, o’, u’ and n’
• Pronunciation:
  – Word into characters (list of phones)
  – No vowel/consonant information available!
  – Each word is coded as a single syllable
  – Numbers?: Expanded into complete words using knowledge base
Labeling

• Typical Process using DTW (in Festvox):
  – Phone set and duration information is available
  – Prompts are generated
  – Use this acoustic and duration information to do a DTW on the uttered sentences

• If no Phone Set?
  – Labeling using acoustic models of speech recognition systems (Sphinx Tools)
  – Acoustic models built using letter as phone names
  – Once the models are trained, segmental information could be obtained.
How Good These Systems are?

• Confirm with different pronunciation of letters in different context
  – Letter context and position information in the word is useful
  – Ex: casa $\rightarrow$ /k a s a/ (house)
  – cesa $\rightarrow$ /th e s a/ (stop)
  – cine $\rightarrow$ /th i n e/ (cinema)
  – cosa $\rightarrow$ /k o s a/ (thing)

• Could capture dialect differences
  
<table>
<thead>
<tr>
<th>Castilian</th>
<th>Colombian</th>
</tr>
</thead>
<tbody>
<tr>
<td>cesa $\rightarrow$ /th e s a/</td>
<td>/s e s a/</td>
</tr>
</tbody>
</table>

• Synthesis Quality: Results show good rating for 90% of words.
Pros and Cons..

- Overcome the effects of using one language/dialect phone set onto another (ex: Pronunciation of Scottish English speaker does not match with US English lexicon!)
- Does not require linguistically knowledgeable speakers of the language
- May not be easy to specify/formulate fine distinctions
- Letter to sound rules may not be easy
- Requires sufficient data for the model to get trained
Speech Recognition

- Pronunciation Dictionary: Core component
- Each lexicon entry is mapped to sequence of sub word units (phonemes)
- Accuracy of ASR systems heavily depend on consistency and accuracy of pronunciation dictionary

- For new languages, automated generation of pronunciation dictionaries rule-based or statistical based approaches.
  - Dictionaries – hand crafting is time consuming
  - Non-accurate dictionaries degrades the ASR performance
Grapheme Vs Phoneme ASR

• Phoneme based systems
  – 3 – state HMM with 3000 triphone models
  – 32 Gaussians for each HMM state
  – Linguistically motivated questions to cluster the polyphonic decision tree

• Grapheme based systems
  – As in the case of phoneme, modeled by 3-state HMM
  – Pronunciation dictionaries: - split the word into characters

• Decision trees for context dependent modeling
## Performance

<table>
<thead>
<tr>
<th>Language</th>
<th>WER</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Grapheme</td>
<td>19.1%</td>
<td>17.0%</td>
<td>26.8%</td>
</tr>
<tr>
<td>Phone</td>
<td>12.7%</td>
<td>17.7%</td>
<td>24.5%</td>
</tr>
<tr>
<td>English</td>
<td>26.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>17.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Grapheme-phoneme Correspondence
Context Width of the Models

- A context width of one (C-1) leads to tri grapheme system
- A hybrid tri grapheme system in which question and model context windows are different.

<table>
<thead>
<tr>
<th>Language</th>
<th>C-1</th>
<th>C-1 Q-2</th>
<th>C-2</th>
<th>C-2 Q-3</th>
<th>C-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>19.1%</td>
<td>19.8%</td>
<td>21.7%</td>
<td>22.4%</td>
<td>23.6%</td>
</tr>
<tr>
<td>German</td>
<td>18.1%</td>
<td>17.0%</td>
<td>18.4%</td>
<td>18.7%</td>
<td>18.7%</td>
</tr>
<tr>
<td>Spanish</td>
<td>27.0%</td>
<td>26.8%</td>
<td>28.8%</td>
<td>28.2%</td>
<td>31.4%</td>
</tr>
</tbody>
</table>
Other Ideas

• Question generation to group poly-grapheme into a limited number of clusters.
• Multilingual Grapheme based Recognition
  – Rapid adaptation to new languages
  – Similar to multilingual phoneme based speech recognition
• Questions and Discussion
• Future directions on the topic.